Building an Index of Nanomedical Resources: An Automatic Approach Based on Text Mining

Stefano Chiesa, Miguel García-Remesal, Guillermo de la Calle, Diana de la Iglesia, Vaida Bankauskaite, and Víctor Maojo

Biomedical Informatics Group, Dep. Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Spain

Abstract. Nanomedicine is an emerging discipline aimed to applying recent developments in nanotechnology to the medical domain. In recent years, there has been an exponential growth of the number of available nanomedical resources. The latter are aimed to different tasks and include databases, nanosensors, implantable materials, etc. This leads to the necessity of creating new methods to automatically organize such resources depending on their provided functionalities. In this paper we will first present a brief overview on the nanomedical discipline and its related technologies. Next we will introduce a method targeted to the automated creation of an index of nanomedical resources. This method is based on an existing approach to automatically build an index of biomedical resources from research papers using text mining techniques. We believe that such an index would be a valuable tool to foster the research on nanomedicine. This is an example of application in the new area of Nanoinformatics.

Keywords: nanomedicine, text mining, biomedicine, nanoinformatics.

1 Introduction

According to the National Nanotechnology Initiative (NNI), "nanotechnology is the understanding and control of matter at dimensions between approximately 1 and 100 nanometers, where unique phenomena enable novel applications" [1]. In 1959, Richard Feynman presented a conference called "There's plenty of room at the bottom"[2]. In his talk, he touched topics that in a few decades became one of the main objectives for the research world. He did not only predict the progressive miniaturizing process that led to the possibility of work at atomic level, he also considered it crucial that the manufacturing of tools can interact with biologic cells at the same size level.

Nanotechnology actually constitutes a path that leads towards the integration of natural and artificial world. One of its strongest points is the fact that nano-particles and nano-devices can effectively interact with natural organism. Considering this context, from the perspective of computer science, nanotechnology catalyzes the passage from its old domain (the study of phenomena surrounding computers) into a more modern one, in which computer science can be defined as the study of natural and artificial information processes [3].

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 50–57, 2008.

The possibility to interact with the natural biological environment makes the research on the biological processes and the understanding of the information processes that are embedded to them easier. Moreover the possibility to manipulate these information flows opens new research opportunities for artificial information processes.

1.1 Nanomedicine Definition

The application of nanotechnology to health care is called nanomedicine. At the end of 2002, The National Institutes of Health (NIH) created a new plan for the study of the nanoscience and nanotechnology applied to the medicine. The European Commission has also shown the same increasing interest in using bio-molecular approaches for the diagnosis, monitoring and treatment of high risk diseases like cancer or cardiovascular conditions and developing micro-nano devices and tools for research and development.

According to Jain [4], nanomedicine is based on three progressively more powerful molecular technologies:

i) Nanoscale-structured materials and devices
ii) Genomics, proteomics and artificially engineered microorganism
iii) Molecular machine systems

These may be used for a large number of application fields that can be organized in the taxonomy [5] can be seen in figure 1.

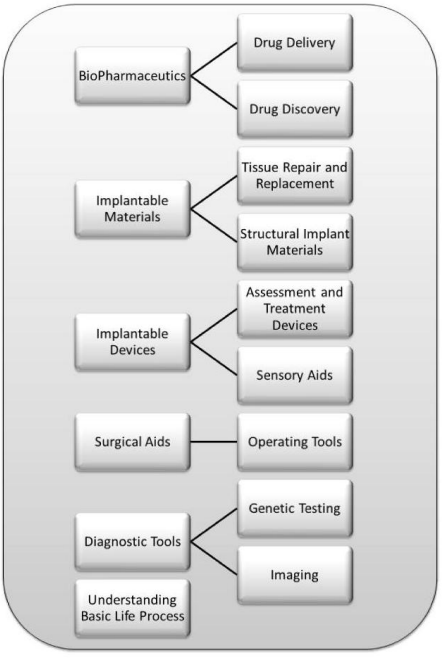

Fig. 1. Nanomedicine taxonomy [5]

1.2 Nanomedicine Taxonomy

Following Figure 1, nanomedicine can be categorized in several different areas:

Biopharmaceutics. This area studies the role of nanotechnology in the pharmaceutical domain. It includes the study of new drugs based on nano-particles (drug discovery) and the nano-systems that can be utilized to deliver the pharmaceutical products in a more effective and precise way (drug delivery).

Implementable materials. This area includes the biocompatible materials that can be permanently or temporally implanted in a living organism. These materials can be used for substituting or repairing tissues (tissue repair and replacement) and structures (structural implant materials) of an organism body.

Implementable devices. This area contains the technologies that aim towards the creation of nano-devices that can be implanted in live organisms. This category comprises those devices that can process local extracted medical information for diagnosing and treating purposes (assessment and treatment devices). Implantable devices also include devices that can enhance sensory skills restoring lost hearing and sigth functions (sensory aids).

Surgical Aids. This area includes the devices that can be helpful for surgical operations. In particular nano tools can be used to perform common surgical tasks in a very precise way or monitoring patient condition with a higher accuracy (operating tools).

Diagnostic tools. This area includes the nano-systems that can help to identify the occurrence of a disease as soon as possible. There are possibilities to work directly on genes and genetic samples (Genetic testing) and to create graphical representations that shows images of the patient's condition (imaging).

Understanding basic life process. Nanotechnology uses devices and tools created at atomic and molecular sizes. For the biological purposes of understanding life process this is a very important opportunity. Through nanotechnology it is possible to deeply understand the processes like protein folding, and to be able to solve problems that are strongly bound within them.

As shown in this taxonomy there are several areas of interest in medicine, which we will analyze below.

2 Background and Rationale

2.1 Nanomedicine Overview

The role of nanomedicine over the coming decades will become progressively more central for patient care process. In the "strategic research agenda for nanomedicine" [6] we can find some research lines that are considered the most valuable in respect of benefits for the patient and socio-economical impact.

Nanomedicine will be used for different medical purposes, such as preventive medicine, diagnosis, therapy and follow up monitoring. Current research covers on these aspects of the care process, with respect to different diseases. Nowadays the most

frequent cause of death in the European Union is cardiovascular diseases, followed by cancer. Other types of disorders, such as musculoskeletal and inflammatory diseases or neurodegenerative diseases, significantly reduce the patient's quality of life.

For these conditions, nanomedicine will offer solutions to reduce the contraindications that current therapies have, whilst being more effective. Current research is rapidly approaching a stage in which nanotechnologies will be utilized in practice. Methods for fields like biological research and biological imaging, applied to medicine, have been developed using nanosize particles and crystals [7].

Nowadays nanomedicine can count on a huge number of new technologies that can be used in the different branches in which nanomedicine is divided [8]. For instance, surfaces perforated with nanopores are used to create containers that can hide the immunologic system biologic cells belonging to other organisms. Anticancer therapies may count on several nanostructures like fullerenes, nanoshells or tectodendrimers that can perform disease recognitions and target delivering tasks.

In a short period (3 to 5 years) nanotechnology is expected to provide biologic robots to medicine, constituted by engineered bacterial organisms able to produce substances useful for the patient metabolism. They may be used, for example, to increase low levels of vitamins or to produce antitoxins when it is needed.

In a longer period (10 to 20 years) nanomedicine will be able to create artificial red and white blood cells or to replace fragments of chromosomes that may cause genetic diseases to the patient.

The research on nanomedicine is growing fast and the available material about this theme is expected to increase dramatically in the next few years.

2.2 Information Management

Over the last few years, the number of papers concerning nanomedicine has increased significantly. The interests on this field has led many research laboratories to increase their effort in obtaining relevant results. The material produced is then shared with the research community.

The rapid growth of produced information leads to organizational issues. Research papers are usually not formally structured. For this reason it is complex to create an automatic approach to collect and order research results. Moreover the possible relation that may exist among different researches is unclear.

This situation may lead to an uncontrolled explosion of hard to manage information. In order to avoid this, automatic approaches to organize the produced material and the existing related concepts (contained in articles) are needed.

Moreover articles may need some standard features in order to structure the information they contain. This may be extremely interesting approach to be able to manage efficiently (automatically) the available resources.

3 Methods

3.1 Automatic Text Analysis

To address the automated creation of a nanomedical index, we propose a method based on automated text analysis using pattern matching techniques. This method has

been previously applied to the biomedical domain to create a tool for automated annotation of biomedical resources from literature [9].

First, linguistic patterns that occur in a text are manually identified using a training set composed of several hundreds of research papers. The extracted patterns describe general syntactic structures and morphological features that allow identification the relevant information to be retrieved from the texts. We considered three different sets of patterns that are used to perform different tasks: i) extracting the names of the resources, ii) identifying their functionalities and iii) classifying the resource into a suitable category.

This classification is driven by the type of resource—e.g. database, annotation tool, visualization tool, etc— and its application domain—e.g. genome, DNA, protein, etc. Types and domains are organized into an ad-hoc taxonomy created by the team of experts.

Using the extracted patterns, we created two transition networks (TN) [10]. These are abstract machines that can determine if a given string belongs to a specific formal language defined by a set of regular expressions.

The first TN extracts the name of the resource and a description of its functionalities, while the second performs the classification task.

The analysis of each document involves 4 stages:

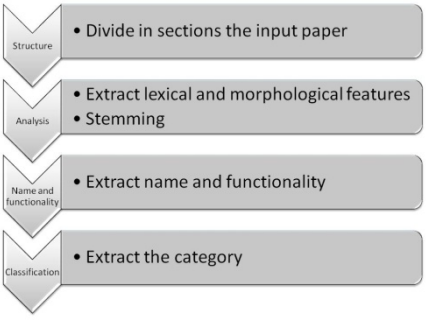

Fig. 2. Text analysis process

Structure. The document is pre-processed to create a surrogate. The latter includes the title, the abstract and a reference to the full text of the article.

Analysis. The title and the abstract sections are divided into single sentences that are parsed using a lexical and morphological analyzer. This produces a sequence of tokens labelled with their corresponding linguistic features. Next, the tokens are stemmed, thus being converted into their root form.

Name and functionality. The tokens are fed as input of the first transition network to extract the name of the resource and a description of its provided functionalities.

Classification. Finally, the tokens are fed as input to the second TN that classifies the tool into the most suitable categories of the taxonomy.

All the extracted information is then stored in a database that can be queried by users through a graphical interface.

3.2 Adaptation to the Nano-context

The results obtained in the biomedical domain, encouraged us to explore other possible application fields. We selected the nanomedical domain due to its growing interest in the biomedical community.

First, we identified the required modifications to be performed on the former approach to adapt it to the nanomedicine field.

Similarly as we did for the biomedical domain, we first extracted linguistic patterns by analyzing abstracts of research papers. These linguistic patterns were then used for the creation of the new TNs tuned for the nanomedicine domain. The first issue we had to tackle was that the number of papers on nanomedical resources was much smaller than those related to biomedical resources. This hindered the patterns' extraction process, making it slower and more complex.

The features we were interested in extracting from the documents were the same as for the previous prototype. We wanted to retrieve the resource's name, its functionalities and the category to which it belonged. For some specific resources we also considered which were their required inputs and outputs –e.g. the output of a nanosensor and the input of a diagnostic nanodevice. This can be used to define complex workflows that may involve a sequence of different resources.

3.3 Enhance the Text Analysis Performance

The implementation of the proposed text analyzer must address a series of problems. Probably the most difficult is solving the heterogeneity introduced by different researchers when describing their results in research papers. This means that we need a more complex set of linguistic patterns in order to be able to cover all the potential possibilities in which a given concept may be expressed.

This leads us to consider the idea of a standard proposal that may help to enhance the performance of text mining tools. The basic idea consists in encouraging authors to prepare the abstracts of their articles following a predefined structure. The latter defines different patterns that facilitate the extraction of relevant information. This may include, for instance, a set of short sentences introduced by a concrete keyword describing the resource presented in the paper.

4 Results and Discussion

The method we described has been tested on the biomedical domain. The set of papers we considered was composed of 392 papers. Among them, a small percentage was related with nanomedicine domain. The first TN managed to recognize, over the whole test set, 376 correct resource names. It also extracts 505 functionalities. The 88% of them were understandable and complete. The 10% were incomplete but still provided useful information. The second TN managed to categorize 305 resources and to assign a domain to 253 resources.

These results refer to the application of the method using TNs mainly tuned for the biomedical field. Our purpose is to refine the set of patterns in order to make them more specific for the nanomedicine domain. The patterns we defined until now are still not stable enough to produce effective results. Nanomedicine is an emerging field and the number of published papers is smaller than in the biomedical domain. High impact journals on nanomedicine are relatively young and is needed a wider set of relevant papers is needed to increase the number of patterns and tune properly our TNs.

5 Conclusion and Future Work

We proposed a new approach to the organization of the information produced about nanomedicine. We focussed on a method that led to previous results in the biomedical domain. We analyzed it in order to identify the characteristics of the field of nanomedicine. This opens a path to the implementation of a "nano-resourceome" [11].

We also proposed a way to produce a structured organization of the documents to improve the efficiency of those tools that have to recover information automatically. In this early stage of nanomedicine research, creating the basis for a common standard of articles is a crucial task to be able to track the rapid growth of information about the domain.

The approach we propose is part of the agenda of the ACTION-grid project. The latter is a project supported by the European Commission beginning in June 2008. The main objective is to constitute an international cooperative action on grid computing and biomedical informatics between the European Union, Latin America, the Western Balkans and North Africa. The project includes the survey of the current state of nanotechnology applied to medicine and the production of a White Paper that highlights future research lines based on the synergy between medical informatics and bioinformatics, expanding results towards grid and nano areas (nanotechnology, nanoinformatics, nanomedicine).

In the context of this project the proposed method is a useful tool for information retrieval about nanomedicine resources. These would be categorized and summarized by name and functionalities, giving the researcher a structured index.

Finally we are working towards implementing a software tool able to find relationships between papers in an automatic way. This enhancement provides the possibility of creating workflows among resources or performs more complex queries on the produced nanoresources' database.

Acknowledgments. This research has been supported by the European Commission-funded ACGT project (IST-2005-026996) and the Spanish Ministry of Education (TSI2006-13021-C02-01).

References

1. National Science and Technology Council.: The national Nanotechnology initiative. Strategic Plan. National Science and Technology Council (2007)
2. Faynman, R.P.: There's plenty of room at the bottom. Eng. Sci. 23, 22–36 (1960)

3. Denning, P.J.: Computing is a Natural Science. Commun. ACM 50(7), 13–18 (2007)
4. Jain, K.K.: The handbook of Nanomedicine. Humana Press, Totowa (2008)
5. Gordon, N., Sagman, U.: Nanomedicine Taxonomy. Canadian Institute of Health Research & Canadian NanoBusiness Alliance (2003)
6. European Technology Platform: Nanomedicine: Nanotechnology for Health. European Technology Platform, Brussels (2006)
7. Alivisatos, A.P.: Less is more in medicine. Sci. Am. 285, 66–73 (2001)
8. Freitas, A.R.: Nanomedicine: Nanotechnology, Biology, and Medicine 1(1), 2–9 (2005)
9. García-Remesal, M., Maojo, V., Crespo, J., Billhardt, H.: Logical Schema Acquisition from Text Based Sources for Structured and non-structured Biomedical Sources Integration. In: Proc. AMIA Symp., pp. 259–263 (2007)
10. Woods, W.: Transition Network Grammars for Natural Language Analysis. Commun. ACM 1970 13(10), 591–606 (1970)
11. Cannata, N., Merelli, E., Altman, R.B.: Time to Organize the Bioinformatics Resourceome. PLoS. Comput. Biol. 1(7), 76 (2005)